US-PAT-NO: 6230153

DOCUMENT-IDENTIFIER: US 6230153 B1

TITLE: Association rule ranker for web site emulation

---------- KWIC ---------

Brief Summary Text - BSTX (3):
The present invention relates to applying data mining association rules to sessionized web server log data. More particularly, the invention enhances data mining rule discovery as applied to log data by reducing large numbers of candidate rules to smaller **rule sets**.

Brief Summary Text - BSTX (5):
Traditionally, discovery of association rules for data mining applications has focused extensively on large databases comprising customer data. For example, association rules have been applied to databases consisting of "basket data"--items purchased by consumers and recorded using a **bar**-code reader--so that the purchasing habits of consumers can be discovered. This type of database analysis allows a retailer to know with some certainty whether a consumer who purchases a first set of items, or "itemset," can be expected to purchase a second itemset at the same time. This information can then be used to create more effective store displays, inventory controls, or marketing advertisements. However, these data mining techniques rely on randomness, that is, that a consumer is not restricted or directed in making a purchasing decision.

Brief Summary Text - BSTX (6):

When applied to traditional data such as conventional consumer tendencies, the association rules used can be order-ranked by their strength and significance to identify interesting rules (i.e. relationships.) But this type of sorting metrics is less applicable to sessionized web site data because site imposed associations exist within the data. Imposed associations may be constraints uniformly imposed on visitors to the web site. For example, to determine a relationship between site pages that web site visitors (visitors) find "interesting" using traditional data mining association rules, a researcher might look at pages that have strong **link** associations. However, for typical web site data, this type of association rule would probably be meaningless because of the site's inherent topology as discussed below.

Detailed Description Text - DETX (17):

Another scenario may arise where page A on a given web site has **links** to pages B, C, and D, and where these three pages are not accessible from **links** off any other page other than A on this site. Then rules B.fwdarw.A, C.fwdarw.A, and D.fwdarw.A may have confidences of 100% for the same reason: namely, traffic flow constraints impose this regularity. On the other hand, consider the rules A.fwdarw.B, A.fwdarw.C, and A.fwdarw.D. Furthermore, assume that page A has no other **links**. If these rules have confidences of 33%, 33%, and 34% respectively, it indicates a very balanced distribution of traffic across these three **links**. This fact might be interesting to the administrator of the site, or even to the web architect whose job it is to arrange the

content on the site to suit the visitors' preferences. On the other hand, it
may be less interesting to those most interested in traffic flowing to page D.
Although it receives slightly more traffic than the other two paces, the
traffic flow it receives is not much more than that which can be explained by
random choice, for example, for visitors making completely random choices at
each decision point.

Detailed Description Text - DETX (18):
Alternatively, it might be of substantial interest if these confidences were
instead 5%, 5%, and 90%. It might be even more interesting where E.fwdarw.D,
where E is a page that does not have direct **links** to either A or D, yet rule
E.fwdarw.D has confidence of 10%. Although 10% may seem like a low number
relative to the examples just considered, this level of confidence may actually
be striking if it is due to apparent strong mutual interest in both E and D
even though the two pages are not directly accessible from each other.

Detailed Description Text - DETX (19):
Currently, eliminating these types of problems by direct analysis of a web
site's topology is either impractical or entirely unachievable. For example,
graph connectivity analysis alone does not suffice, because solving this
problem requires knowledge of the routing between traversal **links**. In actual
web server logs, the situation is complicated by the fact that pages tend to be
accessible in multiple ways, and that **links** can appear on multiple pages.
Furthermore, pages can be created dynamically depending upon the attributes of
the visitor. Because page **content** can determine the **link** traversal topology,
web site topology itself can therefore be dynamic.

Detailed Description Text - DETX (21):

The Web Walker Emulator incorporated by reference above may be used to implement the methods of the present invention. In one embodiment, the Web Walker Emulator is a method for creating a probabilistic generative model of a web site that simulates the behavior of visitors traversing through the site. This simulation "emulates" the behavior of actual visitors to a web site. The parameterization of the simulation can be adjusted in one embodiment such that these "emulated" visitors display behavior that is substantially indistinguishable from those of actual users (or a subset thereof) with respect to population statistics observed over their respective traffic patterns. Or, in another embodiment, it can be tuned to display hypothetical behavior such as visitors acting without evidence of intentional choice. Tracking the site usage traffic of emulated visitors may yield a set of reference distributions ("emulated distributions") against which may be compared the site usage distributions obtained for actual users. The emulated distributions are used to implement estimation methods which measure relative information **content**. The Kullback-Liebler Information Criterion and the Bayesian criteria, widely known to those schooled in the art, are two such estimation methods. The result is a set of reference distributions against which the distributions obtained for actual users may be compared.

Detailed Description Text - DETX (43):

FIG. 1 shows that, through appropriate data access programs and utilities 108, a mining kernel 106 accesses one or more databases 110 and/or flat files (i.e., **text** files) 112 which contain data chronicling

transactions. After executing the steps described below, the mining kernel 106 outputs association rules it discovers to a mining results repository 114, which can be accessed by the client computer 102.

Detailed Description Text - DETX (60):

The ranking method discussed above with respect to FIG. 2 compares the relevance of two sets of rules in which the consequents of the rules comprise a complete set of events, where the relevance of each set is measured by comparing its associational support against the reference given by an emulated distribution. However, rules within the same set may also be compared as shown in FIG. 3. Relative entropy is a measure of "expected" information **content** for discriminating between two distributions--i.e., it is an average value of a pointwise measure. This pointwise measure can be used to compare individual rules within a set of rules. More, precisely: it can be used to compare measures over a set of rules, given that these measures comprise a probability distribution.

Detailed Description Text - DETX (76):

In another embodiment. and under the appropriate conditions (e.g., sufficient data, stationary data generating process), association rules' measures of "confidence" can be used in one method to estimate conditional probabilities. In particular, the confidence of rule A.fwdarw.B gives a useable estimate of the conditional probability P(A.vertline.B). The same techniques as described immediately above may be applied for rule support to compare the confidence of rules against a baseline reference distribution. Relationships such as defined in statements (C) and (D) above

are easily
applied to evaluating rule confidence. With substitution of the appropriate
conditional probabilities, the relationships and the rules
are sorted in steps
706 and 708 where:

Detailed Description Text - DETX (78):
Sorting likelihood ratios as described above are equivalent to traversing
the distribution P.sub.AB /R.sub.AB and looking for places where it deviates
significantly from 1. Peaks (ratios much greater than 1) show where the
confidence of rules under P.sub.AB is significantly greater than what is
suggested by R.sub.AB, and dips (ratios close to 0) show where the confidence
under P.sub.AB is unusually lower than what is suggested by R.sub.AB.
Comparatively speaking, the interpretation of the relationship statement (E) is
not as tidy because the conditional probabilities do not in general lend
themselves to forming a probability distribution; for each i in statement (E)
simply delivers a pointwise measure of the information
content for
discriminating between the two distributions
[P(A.vertline.D.sub.i),P(A.vertline.[character pullout]D.sub.i)] and
[R(A.vertline.D.sub.i),R(A.vertline.[character pullout]D.sub.i)]. The
relationships in statement (F) add the benefit of giving emphasis to rules with
greater support, which is ideally suited to the determining applications for
which these techniques are intended. The method ends in step 710.